图书情报工作 ChinaXiv合作期刊
第 62 卷 第 24 期 2018 年 12 月

多维度属性加权分析的微博用户聚类研究

张海涛[1,2] 唐诗曼[1] 魏明珠[1] 李泽中[1]
[1] 吉林大学管理学院 长春 130022 [2] 吉林大学信息资源研究中心 长春 130022

摘要：[目的/意义]准确把握社交网络用户兴趣倾向，对用户进行分类并形成高聚合的用户群，对研究社交网络信息生态以及信息推荐有重大意义。[方法/过程]通过构造基于多维度的用户属性描述层次模型，根据模型数据需求从新浪微博抓取用户样本数据，对相关用户背景信息、用户博文信息以及用户行为信息的多维度属性下二阶变量进行量化，构造用户向量表达式，比较单一维度与多维度下的用户分类效果，进一步给属性赋予不同的权重值进行加权分析，在取得最优聚类效果后进行方差分析，对模型进行改进。[结果/结论]基于多维度属性加权后的用户聚类效果明显高于单一维度及多维度非加权条件下的用户聚类，且用户博文内容维度对于提高用户聚类效果的有效性最大。

关键词：微博 多维度 用户聚类 加权分析
分类号：G250
DOI:10.13266/j.issn.0252-3116.2018.24.016

ChinaXiv:201903.00453v1

1 引言

近年来，搭载互联网技术迅速崛起的社交媒体不断普及，传统社交模式被打破。搭载互联网技术的社交网络比传统社交网络更为复杂，也存在更多的可研究空间，对其进行更为深入的研究不仅能促进对社会网络的相关研究，更能为社交媒体进一步发展提供指导。

中国互联网络信息中心（CNNIC）发布的第 41 次《中国互联网络发展状况统计报告》数据显示，截至 2017 年 12 月，中国网民规模达到 7.72 亿，互联网普及率为 55.8%；手机网民规模达 7.53 亿，占比达97.5%，移动互联网已渗透到人们生活的方方面面。而微博作为社交媒体，2017 年用户使用率持续增长，达到 40.9%。2017 年新浪 Q3 微博财报数据显示，截至 2017 年 9 月，新浪微博月活跃用户共 3.76 亿，与 2016 年同期相比增长 27%，其中移动端占比达 92%；日活跃用户达到 1.65 亿，较去年同期增长 25%。显然，微博在社交媒体领域占据主导地位，拥有较大影响力。在微博中，用户可以预先给自己添加相关标签，填写好自己的相关信息，如学校、出身年月等，给出用户相关背景信息，而且可以原创微博、转发微博，在使用微博的过程中也产生了关注博主、转发微博、评论微博、点赞微博等信息行为痕迹。

这些信息与信息行为都反映着用户的兴趣倾向，而掌握了用户的兴趣倾向即可在后期通过创建对应的信息推荐模型，进一步提高用户对微博的利用效益，对于微博而言也能更有针对性地进行相关信息推送或相关营销。对用户兴趣进行挖掘，对兴趣相似的用户进行聚类，既能够更好地降低社交网络研究的复杂度，又能够更好地指导个性化信息推荐服务的开展。

2 文献回顾与论文研究思路

2.1 文献回顾

社交媒体兴起于国外，早在新浪微博之前就有了 Twitter、Ins 等用户量巨大的网络社区，国外对于社交网络的研究比我国研究要更早一些。而近些年随着社交媒体用户数量不断高涨，国内相关学者对社交媒体信息传播机制、用户分类、社区发现等相关主题的研究不断增多、不断深入。用户分类是近年来计算机学科与图书情报学科的一个研究热点，计算机学科集中于对聚类算法的研究以及改进，而图书情报学科重在通过用户聚类进行信息个性化推荐，提高信息利用率。图书情报领域从内容、用户行为、用户背景信息等单维度对用户聚类进行的研究较多，但从多维度视角进行

作者简介：张海涛（ORCID：0000-0002-9421-8187），教授，博士生导师，E-mail：zhtinfo@126.com；唐诗曼（ORCID：0000-0002-4355-7963），硕士研究生；魏明珠（ORCID：0000-0001-8430-7461），硕士研究生；李泽中（ORCID：0000-0002-1970-5815），博士研究生。
收稿日期：2018-05-16 修回日期：2018-07-23 本文起止页码：124-133 本文责任编辑：易飞

张海涛，唐诗曼，魏明珠，等. 多维度属性加权分析的微博用户聚类研究[J]. 图书情报工作,2018,62(24):124－133.

chinaXiv:202308.00453v1

的研究却很少。

一些研究从用户行为出发,研究社交媒体用户的分类,主要探讨哪些用户行为特征对用户聚类有效性较大以及聚类算法的优化等。如 M. C. Alarco'n-del-Amo 等[1]根据社交网站的用户使用频率、经验、交互模式等将用户分为"introvert(内向型)""expert-communicator(专业沟通型)""versatile(多才多艺型)""novel(创新型)"四大类,并总结了这 4 类用户的行为特征。张琳等[2]通过基于反映用户信息行为特征的用户粉丝数、关注数、微博数、收藏数 4 个特征变量对微博用户进行聚类分析,并分析每个类别的特征及影响力。

一些学者从内容维度出发,建立用户与内容的信息关联,从而对用户进行分类,如 J. Hannon 等[3]通过计算用户之间基于内容的相关性,分析挖掘用户之间兴趣的相似性,将相似度大用户的进行聚类进而对不同类别进行个性化信息推荐。D. M. Blei 等[4]研究LAD 模型,对用户所发表的信息内容进行主题提取,以类结构进行建模,形成主题文档,每篇文档的主题以概率的形式给出并进行似然估计,解得用户文档相似度,实现文档与用户之间的关联,再根据主题分布进行主题聚类或文本聚类,从而对用户进行聚类。L. J. Hong 等[5]利用两个主题合并使用 LAD 模型的方法构建了一个基于内容的用户聚类新模型。M. Efron[6]设计了一种从多个角度来分析微博内容的方法,为用户信息进行建模,基于内容维度的用户聚类关注用户所表达或者感兴趣的内容在文本上的相似度,将这种文本相似度作为用户相似度,主要关注点是如何建模来更准确地让内容与用户之间形成映射。

除了以上两个维度,也有学者从用户背景信息对用户兴趣、行为表现的影响进行了分析,如徐志明等[7]在用户相似性度量中考虑到了用户背景信息对用户行为的影响,其实证结果表明,用户背景信息对于用户相似性度量具有较大影响力。而基于用户行为或者信息需求两个维度进行用户聚类分析的研究中,常常忽视背景信息对用户聚类分析的影响。

2.2 本文研究思路设计

近几年关于社区用户聚类分析的研究大多基于用户发布的信息内容或者用户行为单维度[8]。但通过文献回顾并结合真实情况分析发现,用户兴趣影响因素往往是多方面的,既与用户背景信息、用户博文内容有关,也与用户信息行为相关,在对用户进行聚类的过程中,应当同时考虑到这 3 个维度对用户兴趣表达的重要性[9]。本文的研究思路如图 1 所示,试图将用户背景信息、用户信息行为、用户博文内容 3 个维度都考虑到,通过收集相关数据,处理数据,得到单维度用户最优聚类、多维度用户最优聚类和加权用户最优聚类,对聚类效果进行对比分析。具体聚类方式是:首先对这 3 个维度属性进行细分,得到影响这 3 个维度属性的二阶变量,从而构造出一个多维度用户属性描述模型;其次,根据该模型获取用户相关维度的数据,并进行数据处理,得到用户向量,采用向量相似性度量用户间的相似度,设定阈值,将基于兴趣特征的用户相似度大于设定值的用户分为一类,而这个相似度是与用户背景信息、用户行为、用户博文内容均相关的;再次,由于这 3 个维度对于用户兴趣表达的重要性不一定是相等的,再根据其对用户描述模型影响强度需要进行加权分析,探求哪种加权条件下的用户聚类效果最好;最后,通过方差分析确定二阶变量中哪些因素对用户描述影响较小,适当地将其刨除,修正用户描述模型,从而取得更好的聚类效果。而如何充分利用用户背景信息、用户博文信息及用户信息行为进行科学分析,如何对这些信息进行定量化处理,如何进行权重分配从而取得最优聚类效果,实现对用户的精准定位,进行兴趣挖掘、信息推荐与精准运营,是本文着重考虑的。

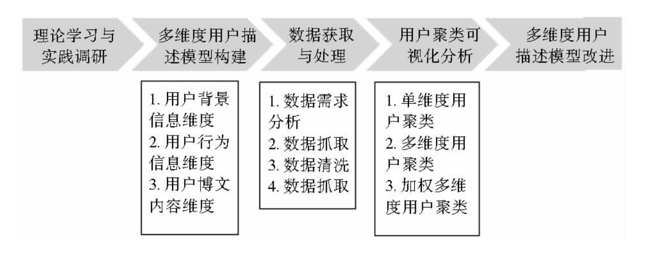


图 1 研究思路

通过参考相关文献[10－11],本文提出了基于用户背景信息、博文信息内容、用户行为信息 3 个维度的用户描述模型,给出的基于多维度用户属性描述模型如图 2 所示:

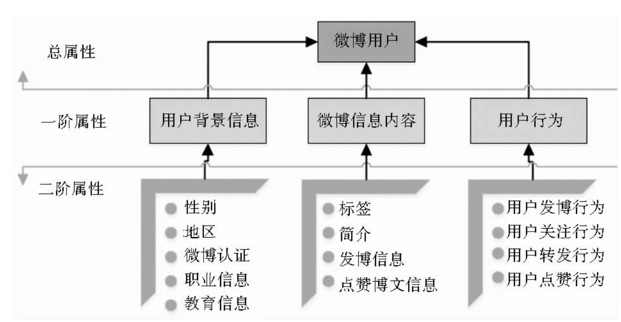


图 2 微博用户多维度属性描述模型

ChinaXiv合作期刊

3 数据获取与处理

3.1 微博数据抓取

对微博用户进行聚类分析首先应该确定哪些特征变量能够较好地反映用户间的差别,以此作为分类基础,以及确定这些特征如何进行量化[12]。微博用户原始特征属性包括粉丝数、关注数、互粉数、个人描述、收藏数、认证情况、性别、注册时间、转发数、话题数、含 URL 数、首条微博发布日、平台数等。很多相关研究在进行数据获取的时候缺少目标导向,所获取的数据有些对于后续研究意义不大,但也予以保留,导致了极大的数据冗余。本文针对前文给出的微博用户多维度属性模型(见图 2),对底层分属性的特征变量进行细分,在借鉴前人相关研究的基础上,得到二阶变量数据需求,如表 1 所示:

表 1 微博用户二阶属性数据需求

标识信息		用户名(作为节点标识,为提高数据处理效率,对用户进行编码存储
用户背景信息(user)		U1 用户性别(采用布尔型数据,1 代表男性,0 代表女性)
		U2 微博认证(采取布尔型数据,1 代表已认证,0 代表未认证)
		U3 地区(采用布尔型数据,1 代表发达城市,0 代表不发达城市)
		U4 教育信息(采用布尔型数据,1 代表有教育信息,0 代表无教育信息)
		U5 职业信息(采用布尔型数据,1 代表有职业信息,0 代表无职业信息)
博文信息内容(content)		I1 标签(提取关键字)
		I2 简介(提取关键字)
		I3 所有博文内容(分词进行词频统计)
		I4 点赞博文内容(分词进行词频统计)
用户行为信息(action)	发博	A1 微博数(采用布尔型数据,1 代表高于平均值,0 代表低于平均值)
	关注	A2 关注数(采用布尔型数据,1 代表高于平均值,0 代表低于平均值)
		A3 粉丝数(采用布尔型数据,1 代表高于平均值,0 代表低于平均值)
	转发	A4 转发博文数(采用布尔型数据,1 代表高于平均值,0 代表低于平均值)
	点赞	A5 点赞博文数(采用布尔型数据,1 代表高于平均值,0 代表低于平均值)

注:表中括号内容为数据处理方式,将在下文数据处理章节详细阐释

本次研究所用到的相关数据均使用数据采集软件八爪鱼采集器进行采集,微博首页将用户兴趣分为时尚、旅游、搞笑、情感、科学、动漫、美食、体育、电影、电视剧、星座、音乐、健身、军事、数码、历史、摄影、萌宠、游戏、美女等 20 个类别,为了形成实验对照组,这 20 个类别暂时予以保留,基于这 20 个类别,在微博找人端口进行相应的类别输入,依照表 1 对每个类别各抓取 120 个用户的相关数据(按照排序依次选取),经过筛选得到最终的用户样本。

3.2 数据处理

3.2.1 数据清洗 对于采集到的 20 个类别各 120 个微博用户数据进行数据清洗,主要是为了避免数据稀疏性以及冷启动问题对后续数据分析以及用户聚类效果产生不良影响。在进行数据清洗的过程中,不仅需要注意将简介、介绍、标签等数据项内容为 NULL 值的用户样本数据进行删除,对于简介或者介绍内容无实际意义的对应项(如"工作联系:XXXXXXX""心若止水")也进行了删除。同时,考虑到到用户主体为机构的微博用户,其在微博中进行的各种活动代表机构而非个人,针对性较强,可研性较弱,故将微博机构认证的微博用户予以删除。而且由于之后要针对用户样本进行各自最近的 20 条博文采集并进行文本分析,将博文数少于 20 的微博用户也进行删除。在剩余的样本中随机抽取 20 * 50 的样本量作为最终用户样本。

3.2.2 基于云模型的文本信息量化 基于微博信息内容的用户分类通常通过构建用户 - 文本对照模型,例如 LDA 模型、三层贝叶斯模型,通过计算结构化文本相似性而计算用户间的相似性,从而进行用户聚类。但是考虑到文本描述模型与用户多维度属性描述模型的不兼容,在进行数据处理的过程中,采用云模型将定性的用户博文信息进行定量表示[13]。

将用户节点视为云模型中的云滴,将原定的 20 个类别视为用户评分的 20 个评分项目。对用户微博信息内容进行处理,通过词袋模型,忽略掉文本的语法、语序,仅将其看作是若干个词汇的集合,文档中每个单词的出现都是独立的,使用分词工具 ik-analayer 对微博信息文本进行分词并进行词频统计,根据词频对项目进行评分,得到用户评分表(样例见图 3)。用户名一栏中为目标用户的微博帐户名称,在用户评分表中统计目标用户的已评分项目集合 S1。S1 为所有项目组成的集合,计算用户的为评分项目 S2 = S-S1,S 是所有项目组成的集合。根据用户评分表可以得到用户 - 项目矩阵,根据矩阵得到用户 i 与用户 j 之间的相似度 Sim(i,j)。这是基于用户博文信息内容单维度的数据处理办法,在考虑用户背景信息、用户博文信息、用户行为这 3 个维度同时作用的情况下,该用户 - 项目的矩阵移植性强,对于用户聚类而言,只需要将用户行为

chinaXiv:202308.00453v1

与用户背景信息的二阶变量作为项目评分表的新增项目,即可简化操作,直接构建用户－项目矩阵,通过同样的操作进行用户相似度的计算,从而进行后续的聚类操作。

用户名	时尚	旅游	搞笑	情感	科学	动漫	美食	体育	电影	电视剧	星座	音乐	健身	军事	数码	历史	摄影	萌宠	游戏	美女
Seoul韩流	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
TVB剧评社	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0
william_彭	1	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1
爆料王	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
本仙女爱追剧	0	2	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0
穿帮君	0	0	0	0	0	0	0	0	2	3	0	0	0	0	0	0	0	0	0	0
大家字幕组	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	3	0	0	0
电视剧透社	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
电视剧周刊	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	1	0
电视圈大哥	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
毒舌八卦	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
樊少皇	0	0	0	0	0	0	0	0	4	2	0	0	0	0	0	0	0	0	0	0
凤凰天使剧讠	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
港剧猫	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
高希希	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0
关晓彤	0	0	0	0	0	0	0	0	2	3	0	1	1	0	0	0	1	0	0	0
好剧台词	1	1	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0

图 3　基于博文信息内容的用户评分表(部分截图)

3.2.3　数据标准化　数据标准化的好处就在于可以提高精度,对于基于距离计算的用户相似度测量算法效果显著,标准化可以让各个特征变量对结果作出的贡献相同。在多维度描述模型中,由于各个维度的性质不同,通常具有不同的量纲和数量级,当各维度间的水平相差很大时,直接用原始指标进行分析,就会突出数值较高的指标在综合分析中的作用,相对削弱数值水平较低指标的作用。因此,为了保证结果的可靠性,需要对原始指标数据进行标准化处理。本研究首先考虑基于用户行为和基于用户背景信息的数据二值化,具体处理办法已在表 1 中给出。

4　数据分析与讨论

4.1　单维度用户聚类可视化分析

本次聚类使用 Tableau 10.5 软件,分别对用户行为数据、用户自身信息数据、博文内容数据进行分析,再基于这 3 个维度同时对这 3 项数据进行非加权分析,得到不同的聚类结果,对这 4 种情况下的聚类效果进行纵向比对分析,并且将这 4 种聚类效果与原本的 20 个类别进行横向比对分析,为了方便横向比对,在进行聚类可视化的过程中选择关键词即用户原本所属类别作为横轴,聚类后的群集作为纵轴[14]。

Tableau 10.5 使用 k 均值算法进行群集。对于给定的聚类簇数 k,算法将数据划分为 k 个类。每个类都有一个中心(质心),它是该类中所有点的平均值。通过 K 均值迭代过程来查找中心,该过程可最大程度地缩短类中各个点与类中心之间的距离。Tableau 将 Lloyd 算法与平方欧氏距离结合使用来计算每个 k 的 k 均值聚类。与拆分过程结合使用来确定每个 k＞1 的初始中心,生成的聚类是确定性的,结果仅取决于聚类簇数。

使用 Calinski-Harabasz 标准来评估聚类质量从而确定最佳聚类簇数。Calinski Harabasz 标准的定义如式(1)所示:

$$\frac{SS_B}{SS_W} \times \frac{(N\text{-}k)}{(k\text{-}1)} \qquad \text{式(1)}$$

其中 SS_B 是类间总体方差,SS_W 是类内总体方差,k 是聚类簇数,N 是观察次数。此比率的值越大,类的内聚性越高(群集内方差小)并且单个类的离散性/分离性也越高(类间方差大)。在确定最优聚类簇数的时候将选择与第一个局部 Calinski-Harabasz 指数最大值对应的簇数。

首先对用户行为数据进行聚类分析。当类的个数为 5 时,取得最优的聚类效果,聚类可视化结果见图 4。在用户行为单维度聚类中,关键词为情感的用户单独被分为一类,类间属性表现为关注数多、微博数多、点赞数多、粉丝数少,转发数少;关键词为"电视剧""电影""动漫""星座""游戏""音乐"的用户被分成一类,类间属性表现为关注数少、发博数少、点赞数中等、粉丝数中等、转发数中等;关键词为"动漫""旅游""时尚"的用户被分为一类,类间属性表现为关注数中等、微博数偏少、点赞数少、粉丝数多、转发数中等;关键词为"健身""军事""科学""历史""萌宠""体育"的用户被分为一类,类间属性表现为关注数中等、发博数中等、点赞数多、粉丝数中等、转发数多;关键词为"美女""美食""摄影""数码"的用户被分为一类,类间属性表现为关注数较多、发博数中等、点赞数偏少、粉丝数中等、转发数中等。

chinaXiv:202308.00153v1

ChinaXiv合作期刊

该维度下，原本的20类用户被高聚合为5类，类间的区别度较高，但是维度属性对用户总属性的影响较小，仅从这单一维度进行用户聚类显然是十分不准确的。

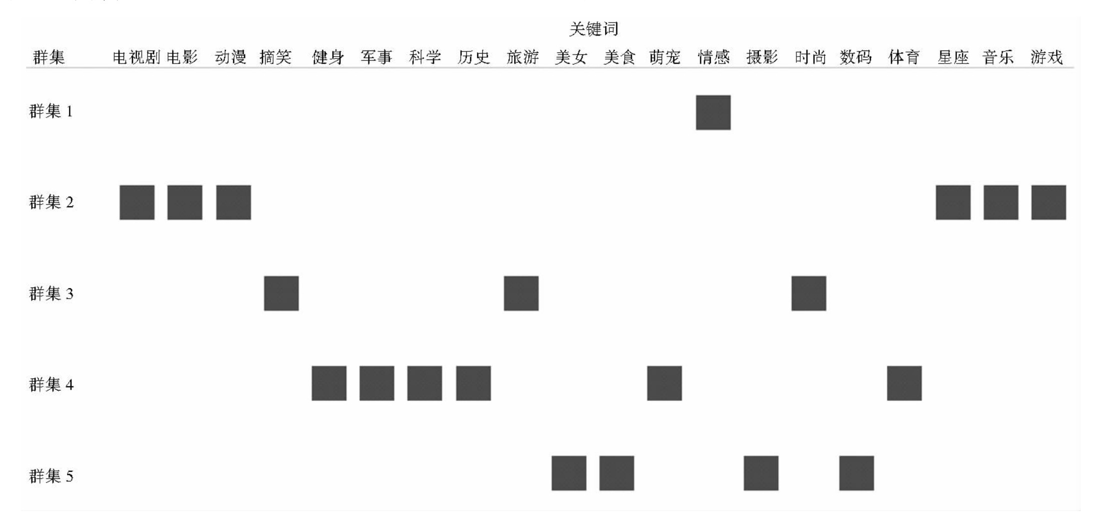


图4 基于用户行为信息单维度的最优用户聚类可视化

基于用户背景信息单维度聚类结果如图5所示，当类的个数为4时取得最优聚类效果。关键词为“军事”“美食”“情感”“音乐”的用户被分为一类，类间属性表现为地区界定不明显、微博认证用户占比大、性别分布均匀、多给出自身教育信息与职业信息；关键词为“电视剧”“电影”“动漫”“历史”“旅游”“星座”的用户被分为一类，类间属性表现为地区分布均匀、微博认证用户占比大、性别分布均匀、给出自身教育信息的用户占比中等但给出自身职业信息的用户占比小；关键词为“搞笑”“健身”“科学”“萌宠”“摄影”“数码”“游戏”的用户被分为一类，类间属性表现为地区分布偏向于非发达城市、微博认证用户占比中等、男性用户占比大，给出自身教育信息和职业信息的用户占比均偏小；关键词为“美女”“时尚”“体育”的用户被分为一类，类间属性表现为发达城市用户占比大、微博认证用户占比小、女性用户占比大、给出教育信息的用户占比小，给出职业信息的用户占比偏小。

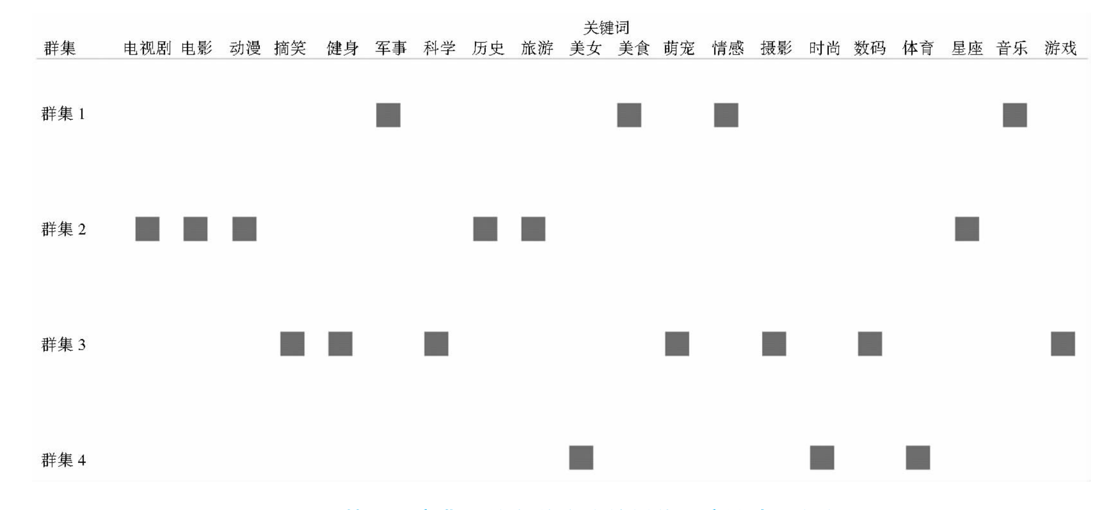


图5 基于用户背景信息单维度的最优用户聚类可视化

从聚类结果不难看出，基于用户背景信息单维度对用户进行分类有一定的借鉴性，其借鉴性比基于用户行为单维度的聚类方式来说信度更高，例如女性更关注美女、时尚等领域，男性更关注健身、科学、摄影、数码、游戏等领域，发达城市的用户隐私保护意识更强，更不愿意给出自身的教育信息、职业信息等，都能从该维度下的用户聚类结果中反映出来，类间间距较大，但基于用户背景信息单一维度的用户聚类的内聚性不够强，仅使用该维度对用户进行分类对于微博内容聚合缺乏指导性。

chinaXiv:202308.00453v1

基于用户博文信息单维度聚类结果如图6所示，当类的个数为2时取得最优聚类效果。关键词为搞笑、情感、星座的用户被分为一类，类间属性表现为词频集中且频繁出现于分属性I1标签、I2简介、I3所有博文内容、I4点赞博文内容这4项中，同时"搞笑""情感""星座"这三大关键词与其他17个关键词共现频数高。剩下的关键词除"搞笑""情感""星座"之外的用户被分为一类，类间属性表现为词频分散且关键词与4项分属性的共现频数低，与"搞笑""情感""星座"这3个关键词的共现频数也低。

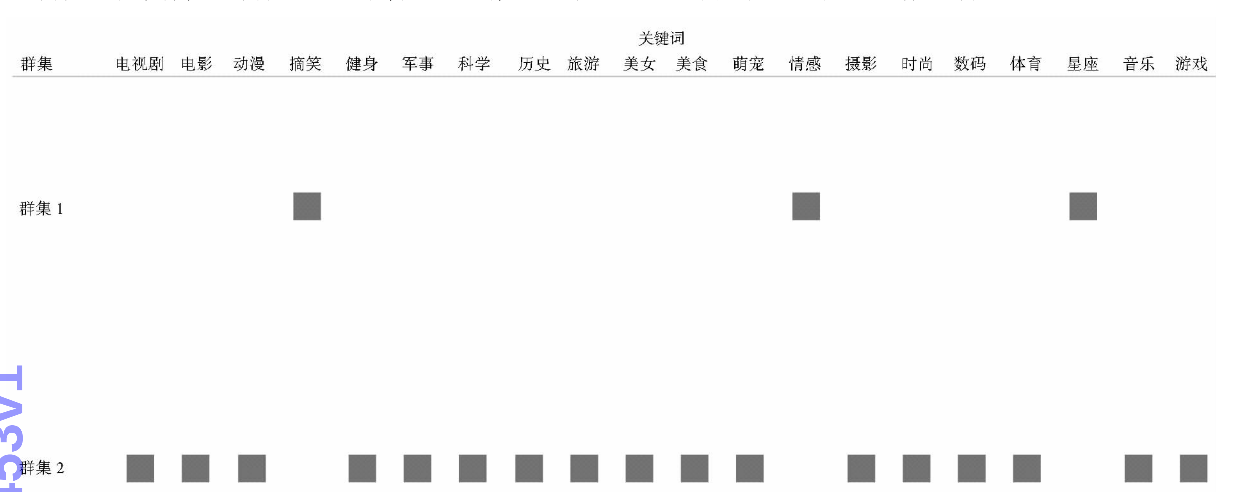


图6 基于用户博文信息单维度的最优用户聚类可视化

基于用户博文信息单维度下的用户聚类在非加权条件下的最优解明显存在重复计算带来的巨大分类误差，聚类的目标在于将博文内容相似度高的用户分为一类，这种相似度高低体现在纵向的用户之间在一阶变量上所体现出来的倾向性，但是在没有对数据进行加权分析时，当用户同时在多项二阶变量上有取值时，用户评分表中对各项评分所进行的运算是简单地累加，对于用户向量来说，这样的运算方式使得向量相似性度量失真。在实例中的表现为：如果用户的标签、简介、博文、点赞博文中重复出现与关键词"搞笑"相关的词语如"段子""笑话""恶搞"等，但与其他关键词相关的词语较少出现，则在现有的评分表运算过程中自然给这种倾向性自动赋予了较高的权值。当一个用户在标签、简介、博文、点赞博文中出现了横跨两个及两个以上的关键词的倾向性时，根据现有的运算规则，其被自动赋予的权重是肯定会比前一种情况要低的，但是前一种情况只能说明该用户的博文内容倾向于关键词"搞笑"，也就是说这样的用户可能只对"搞笑"类信息感兴趣，后一种情况中的用户可能既对"搞笑"类信息感兴趣也其他某类信息感兴趣，但是这并不能说明该用户对于"搞笑"类信息和其他某类的信息感兴趣的程度不如前者。

基于用户背景信息和用户行为信息的用户聚类所用数据源是高度标准化的二值矩阵，因此基于用户博文信息单维度数据同样需要进行标准化，既是为了保证聚类效果比对分析的准确性，也是为了保证后续加权分析数据的可用性。

基于用户博文信息的标准化使用数据分析软件SPSS内置的Z-score即正规化方法，Z-score基于原始数据的均值(mean)和标准差(standard deviation)进行数据的标准化。将原始值X使用Z-score标准化到X'。数据标准化后，基于用户博文信息单维度的用户最优聚类结果见图7。

基于标准化用户博文信息单维度的用户聚类，在类的个数为6时，取得最优聚类效果。从图7可以看出，在一定误差范围内，用户博文信息中表现的用户兴趣倾向往往表现为多个方面，关键词为"搞笑""美女""情感""星座"的用户被分为一类，类间属性表现为简介与点赞内容相似度高，标签与发博内容相似度低。关键词为"动漫""科学""萌宠""数码""音乐""游戏"的用户被分为一类，类间属性表现为用户标签、点赞内容相似度较高，简介、发博内容相似度低。关键词为电视剧和电影的用户被分为一类，类间属性表现为简介、标签、发博内容、点赞内容均偏高。关键词为"旅游""美食""摄影""时尚"的用户被分为一类，类间属性表现为用户标签、发博内容、点赞内容相似度高，简介相似度低。关键词为"军事""历史"的用户被分为一类，类间属性表现为简介相似度低，发博内容相似度中等，标签与点赞内容相似度高。关键词为"健身"和"体育"的用户被分为了一类，类间属性表现为用户发博内

chinaXiv:202308.00453v1

ChinaXiv合作期刊

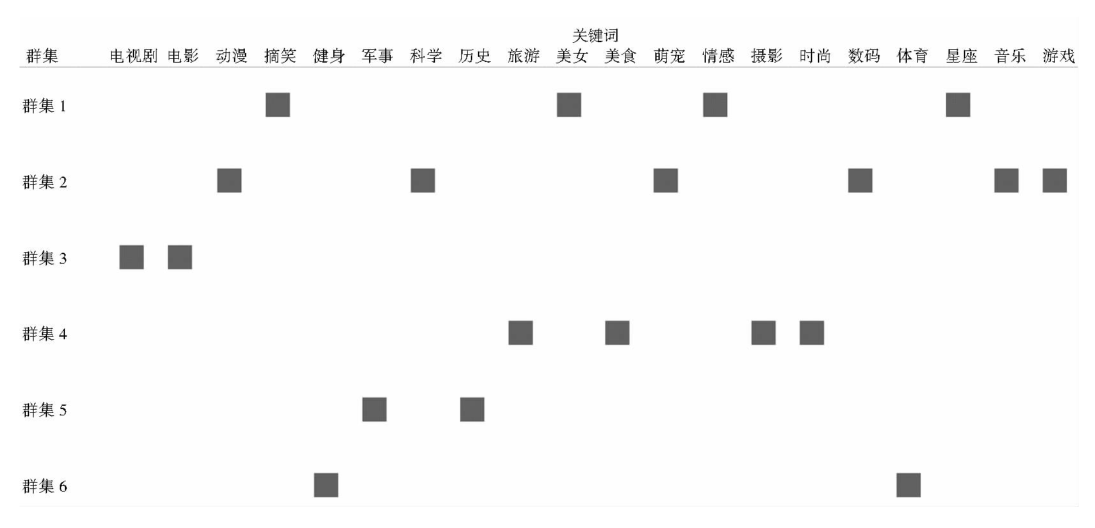


图 7　基于标准化用户博文信息单维度的用户最优聚类可视化

容、点赞内容相似度高，标签内容相似度中等，简介内容相似度低。

标准化后的数据明显提高了基于用户博文信息单维度用户聚类模型的有效性。聚类结果的可借鉴性高，例如，关键词为“电影”和“电视剧”的用户被分为一类，且类间属性具有高相似性，这反映了对电视剧感兴趣的用户往往也表现出对电影类信息的兴趣与关注。

4.2　基于多维度属性加权的用户聚类可视化分析

4.2.1　多维度非加权用户聚类可视化分析　数据经过标准化处理后，得到基于云模型的总用户评分表，将总用户评分表导入软件 Tableau 进行用户聚类分析，得到最优用户聚类如图 8 所示：

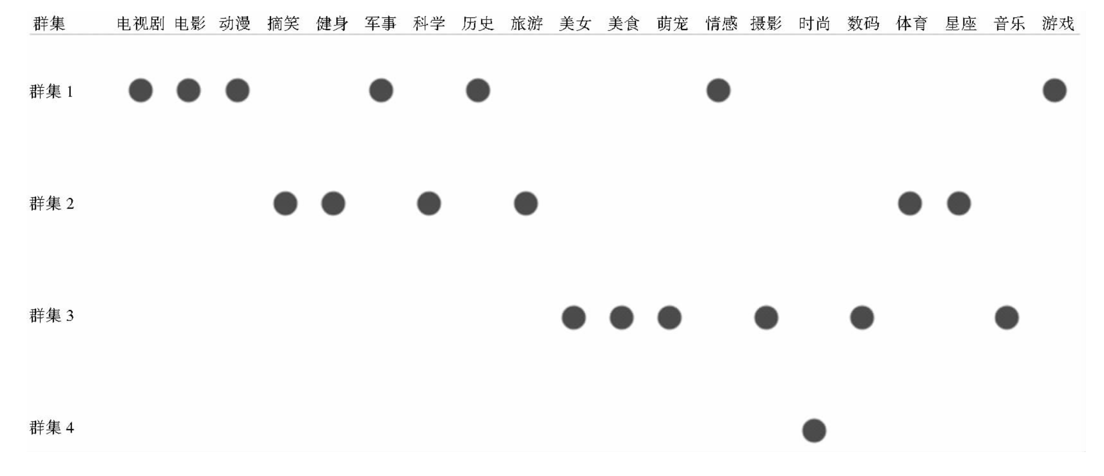


图 8　基于多维度的用户最优聚类可视化

在类的个数为 4 时，得到了最优用户聚类，聚类汇总诊断得到组间平方值总和为 9.182 1，组内平方值总和为 8.219 7。组间平方和指标将每个类之间的间隔量化成为每个类的中心与数据集中心之间的平方距离总和，类中心度量采用平均值，通过分配给类的数据点数进行加权得到。组间平方和值越大，类之间的间隔就越好。组内平方和指标将类的内聚性进行量化，量化为每个类中心与类中单个标记之间的平方距离总和，组内平方距离总和越小，群集的内聚性就越高。分析得到基于多维度的用户聚类在最优情况下聚类效果仍然不是很好，多维度的用户聚类虽然更全面地考虑到了各个维度对于用户相似性度量的影响，但是却分散了用户相似度在各个维度的表现强度，导致聚类效果不够好，组间距离较小而组内距离较大，类别界限不够清晰。

4.2.2　多维度加权用户聚类可视化分析　考虑到 3 个维度在用户聚类过程中可能存在不等贡献，对模型中的每个维度赋予不同的权重，由于有 3 个维度且在之前缺乏相关权重分配研究，在进行加权分析的同时需要进行权重分配试验，本文采取常用的线性加权方

chinaXiv:202308.00045v1

法对3个维度进行加权试验,权重分配如下所示:

加权1:用户向量 = (User * 0.25, Content * 0.5, Action * 0.75)

加权2:用户向量 = (User * 0.25, Content * 0.75, Action * 0.5)

加权3:用户向量 = (User * 0.5, Content * 0.25, Action * 0.75)

加权4:用户向量 = (User * 0.5, Content * 0.75, Action * 0.25)

加权5:用户向量 = (User * 0.75, Content * 0.5, Action * 0.25)

加权6:用户向量 = (User * 0.75, Content * 0.25, Action * 0.5)

在这6种加权方式下,不同加权下的用户最优聚类诊断数据如表2所示:

表2　基于多维度加权的最优用户聚类效果诊断

指标＼加权	加权1	加权2	加权3	加权4	加权5	加权6
聚类簇数	3	6	5	14	8	3
组间平方值总和	5.627 9	10.799	9.848 3	1.643 8	8.156	6.466 1
组内平方值总和	15.06	5.602 8	7.553 5	18.5	10.132	10.936

比对表2中6种加权方式下的最优用户聚类的聚类效果,加权4下的用户聚类效果最好,即当用户博文内容所占权重为1/2,用户背景信息所占权重为1/3,用户行为信息所占权重为1/4,且聚类簇数为14时,取得最优聚类效果,其聚类可视化如图9所示:

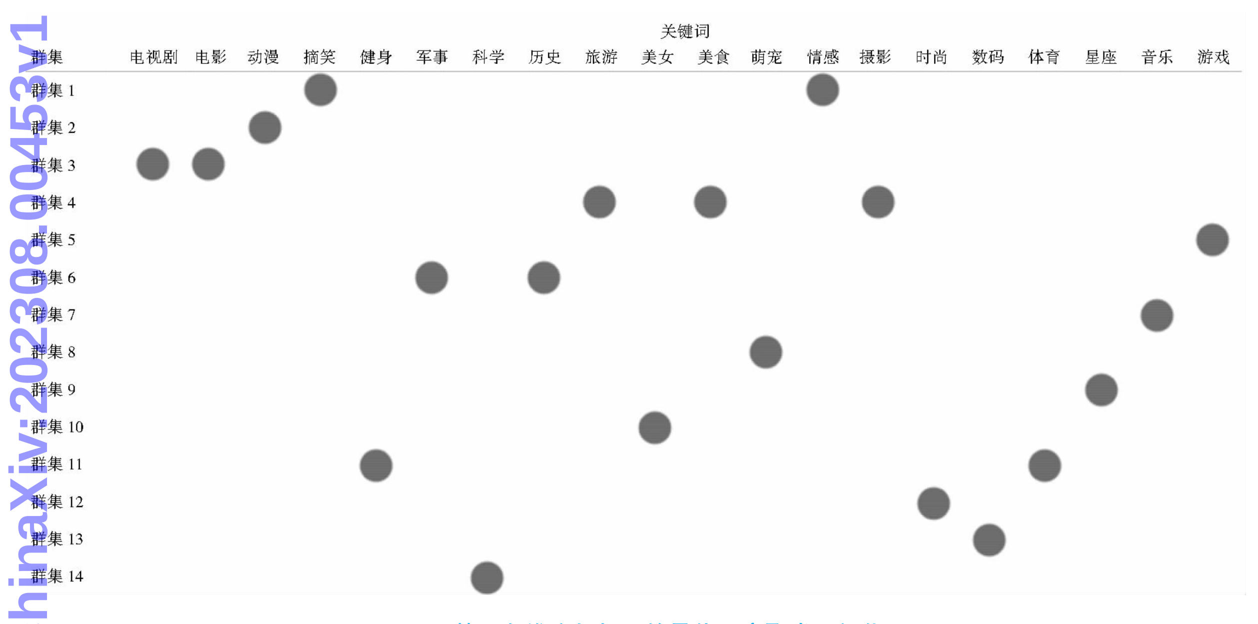


图9　基于多维度加权4的最优用户聚类可视化

该方法与基于用户博文信息单维度、基于用户背景信息单维度、基于用户行为信息单维度的用户聚类效果进行横向比对,类间的特征属性更明确,例如在基于用户博文信息单维度的用户最优聚类中,关键词为"搞笑""情感""美女""星座"的用户被分为一类,但是在多维度加权条件下,关键词为"搞笑""情感"的用户被分为一类,但是关键词为"美女"的用户和关键词为"星座"的用户被单独分成了两个类别,这说明了基于用户背景信息和基于用户行为信息两个维度的相似度差异给聚类结果带来了影响。同理,造成聚类结果与基于单维度的用户聚类结果不同的是另两个维度带来的影响。这种影响增大了用户差异,使得聚类结果更能准确地反映用户兴趣倾向等。

与基于多维度的非加权用户聚类效果进行纵向比对,加权后的用户聚类更符合实际情况。例如非加权情况下,关键词为"电视剧""电影""动漫""军事""历史""情感""动漫"的用户被分为了一类,在加权条件下,分类有了较大幅度的变化。在将多维度进行平等对待的情况下,就可能出现分类异化,可能用户博文相似度较大的用户由于在用户行为和用户背景信息两个维度表现出来的相似度较低而被分散至多个类别,而使实际情况中,基于用户博文内容、基于用户背景信息和基于用户行为信息这3个维度下的用户相似性对于用户间总的相似性度量做功是不同的,只有重视这种差异性才能得到更仿真的用户相似性度量模型和更符合实际情况的用户聚类结果。而通过加权分析证实了在基于用户博文内容单维度所占比重最大的情况下取得的聚类效果是最优的,说明用户博文内容单维度对用户相似性度量的影响力最大。

chinaXiv:202308.00453v1

ChinaXiv合作期刊

同时,与微博原先的分类标准相比,一些原本不属于一个类别的用户被合成了一个类别,例如关键词为"搞笑""情感"的用户被分为了一类,这说明了在进行用户分类的过程中,必须重视重叠社区这一概念,重视用户兴趣间的交叉,考虑非线性用户分类的重要性,而在进行后期的信息推荐时,应当综合考虑用户分类的重叠性,对于有重叠兴趣的用户适当推送其所属类别外的其他类别信息。

4.3 模型优化

4.3.1 *方差分析* 方差分析(ANOVA)是统计模型及关联程序的集合,用户分析已区分为类的观察值内和观察值之间的差值,将每个变量计算方差进行分析,生成的方差分析表可用于确定对群集最有效的变量。Tableau 群集的相关方差分析统计数据包括 F 统计数据、P 值、模型均方值与误差平方和。F 统计数据,单向或者单因素 ANOVA 的 F 统计数据是变量所解释的方差分数,它是组间方差与总方差的比率,F 统计数据越大,在群集之间就能更好地区分对应变量。P 值是指 F 统计数据所有可能值的 F 分布的值大于变量实际 F 统计数据的概率,如果 P 值低于指定的显著性水平,则可以拒绝零假设(变量的单独元素是单个群体的随机样本)。此 F 分布的自由度为(K-1,N-K,其中 K 是类的个数,N 是已建立类的项数)。P 值越低,对应变量的元素的预期值在类之间的区别越大。

模型均方值是组间平方和与模型自由度的比率。组间平方和是对群集均值之间差值的度量。如果群集均值彼此很接近(因此与总均值也很接近),则值将很小。模型的自由度为 k-1,其中 k 为群集数。误差平方和是组内平均和与误差自由度的比率。组内平方和测量每个群集内的观察值之间的差值。误差的自由度为 N-k,其中 N 是已建立群集的总观察值数(行数),k 为群集数。可以将误差平方和看作是总体均方误差,并假定每个群集中心都表示每个群集的"真实值"。在最优加权条件下,模型的方差分析结果如表 3 所示:

表 3 最优加权聚类方差分析

变量	F - 统计数据	p 值	模型		错误	
			平方值总计	DF	平方值总计	DF
所有博文内容 标准偏差	12.922	0.001 323	1.295	13	1.59	7
标签 标准偏差	9.164	0.002 703	1.529	13	2.109	7
微博数 标准偏差	7.931	0.004 951	0.926 6	13	2.004	7
性别 标准偏差	7.115	0.005 113	0.283 7	13	1.14	7
点赞博文内容 标准偏差	5.999	0.016 61	0.240 6	13	1.023	7
关注数 标准偏差	2.575	0.043 58	0.372	13	2.008	7
微博认证 标准偏差	1.03	0.082 82	0.138 1	13	1.14	7
地区 标准偏差	1.018	0.097 24	0.142	13	1.186	7
简介 标准偏差	0.946 7	0.102 5	0.187 3	13	1.681	7
粉丝数 标准偏差	0.824 6	0.255 2	0.176 5	13	1.82	7
转发数 标准偏差	0.822	0.2636	0.113 6	13	1.175	7
职业信息 标准偏差	0.576	0.4759	0.106 8	13	1.17	7
点赞数 标准偏差	0.288 3	0.6662	0.070 98	13	1.026	7
教育信息 标准偏差	0.241	0.788 5	0.045 86	13	1.617	7

由表 3 可知在 14 项二阶变量中,对于模型贡献量最大的是发博内容这一变量,P 值小于 0.1 的变量有 8 项,分别是用户博文信息维度的发博内容、标签、点赞博文内容;用户背景信息维度的性别、微博认证、地区和用户行为信息的微博数和关注数,也就是说在原先的用户描述模型中,这 8 个二阶变量对于用户描述来说具有概括性。其实践意义在于可以分辨出二阶变量对于用户向量表达的影响力力度大小,F 值越大,P 值越小,表示其用于用户聚类的作用力越可靠,也就是说这个变量的存在越能区别不同类的用户。例如,二阶变量发博内容的 F 值为 12.922,P 值为 0.001 323,就说明有 1 - 0.001 323 即 99.867 7% 的把握认为用该变量作为区分用户的变量的结论是正确的。而统计学中一般以 0.01 作为 P 值的标准,P 值小于 0.01 即可认为该变量对于模型是完全有效的。而对于表 4 中 P 值大于 0.01 的二阶变量,在用户描述模型中可以考虑删去以提高模型有效性。

4.3.2 *用户多维度属性描述模型改进* 在最优的加权条件下,通过对最优用户聚类进行方差分析,考虑到二阶变量简介的 P 值十分接近于 0.01,以 P = 0.011 为标准界限,对于 P 值小于 0.11 的二阶变量予以保留,对于 P 值大于 0.11 的二阶变量予以删除,在考虑到最优加权,得到改进后的基于多维度属性加权的用户描述模型见图 10。

5 结语

通过对用户样本数据的聚类分析、加权分析以及

chinaXiv:202308.00453v1

ChinaXiv合作期刊

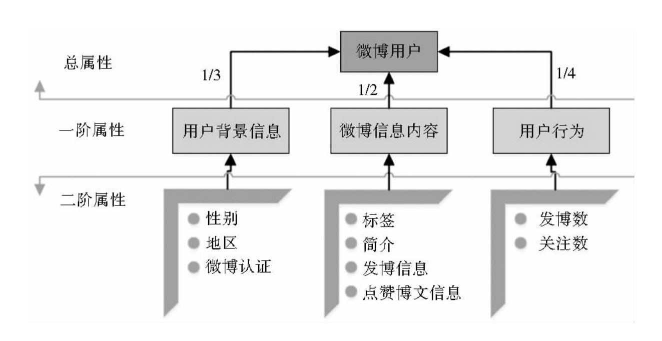


图 10　多维度加权用户属性描述模型(改进后)

方差分析,研究基于多维度属性和加权下的用户聚类效果,得到了基于多维度加权的用户属性描述模型,该模型对于指导微博用户分类以及后续的信息推荐研究意义重大,同时为社区用户聚类提供了新思路。本文还存在着一些不足,如:基于加权思想的用户聚类在不同聚类方法下的聚类效果是否有差异？如何基于该模型进行信息推荐机制的构建？这些还需要进行后续研究。

参考文献:

[1] ALARCóN－DEL-AMO M C, LORENZO-ROMERO C, GóMEZ-BORJA M Á. Classifying and profiling social networking site users: a latent segmentation approach[J]. Cyberpsychology, behavior, and social networking,2011,14(9): 547－553.
[2] 张琳,谢忠红. 基于聚类的微博用户类型与影响力研究[J]. 情报科学,2016,34(8):57－61.
[3] HANNON J,BENNETT M,SMYTH B. Recommending Twitter users to follow using content and collaborative filtering approaches[C]// Proceedings of the 4th ACM conference on pecommender systems. New York:ACM,2010:199－206.
[4] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of machine learning research, 2003,3(1):993－1022.
[5] HONG L J,DAVISION B. Empirical study of topic modeling in Twitter[C] //Proceedings of the first workshop on social media analytics. New York:ACM Press,2010:80－88.
[6] EFRON M. Information search and retrieval in microblogs[J]. Journal of the American Society for Information Science and Technology,2011,62(6): 996－1008.
[7] 徐志明,李栋,刘挺,等. 微博用户的相似性度量及其应用[J]. 计算机学报,2014,37(1):207－218.
[8] 黄静. 消费型虚拟社区的用户行为特征及其应用研究[J]. 图书情报工作,2011,55(3):97－100,51.
[9] 崔金栋,孙遥遥,王欣,等. 基于 Folksonmy 和本体融合的微博信息推荐方法研究[J]. 情报科学,2015,33(10):27－31.
[10] 薛云霞. 微博用户属性识别方法研究[D]. 苏州:苏州大学,2015.
[11] 顾晓雪,章成志. 标注内容与用户属性结合的标签聚类研究[J]. 现代图书情报技术,2015(10):30－39.
[12] 彭希羡,朱庆华,刘璇. 微博客用户特征分析及分类研究——以“新浪微博”为例[J]. 情报科学,2015,33(1):69－75.
[13] 张国英,沙云,刘旭红,等. 高维云模型及其在多属性评价中的应用[J]. 北京理工大学学报,2004(12):1065－1069.
[14] 李小慧. 基于 Jaccard 项目类别相似性的个性化推荐算法研究[D]. 长沙:中南大学,2010.

作者贡献说明:

张海涛:负责文章结构框架设计指导;
唐诗曼:负责论文撰写与修改;
魏明珠:负责数据抓取相关工作;
李泽中:负责方差分析。

Research on the Clustering of Microblog Users Based on Multi-dimensional Attribute Weighting Analysis

Zhang Haitao[1,2]　Tang Shiman[1]　Wei Mingzhu[1]　Li Zezhong[1]
[1] The Management College of Jilin University,Changchun 130022
[2] The Information Resource Research Center of Jilin University,Changchun 130022

Abstract: [**Purpose/significance**] It is of great significance for the study of social network information ecology and information recommendation to accurately grasp the interest tendency of social network users and classify users into highly aggregated user groups. [**Method/process**] In this paper, by constructing the user attributes describe hierarchical model based on multi-dimensional, according to the model data requirements fetching user sample data from Sina microblog, quantify the secondorder variable based on the multi-dimensional property of the users' background information, users' blog information and user behavior information to construct user vector expression, comparing the classification results based on single dimension and the multi-dimensional, given different weights to attribute for weighted analysis, when achieve the optimal clustering results, based it do variance analysis to improve the model. [**Result/conclusion**] User clustering effect based on the multi-dimensional attribute weighting is significantly better than the user clustering effect based on the single-dimensional and under the condition of the multidimensional unweighted, and users microblog content dimension for improving the validity of user clustering effect is the largest.

Keywords: microblogs　multi-dimensional　user-cluster　weighted-analysis

chinaXiv:202308.00453v1